TalkBank and CLARIN

Brian MacWhinney
CMU - Psychology
send to clarin@clarin.eu by
Oct 21 — 15+10 minutes
Basic Questions

- How did language emerge in the species?
- How does it change?
- How is it learned?
- How is it processed?
- What are the results of damage and variation?
## Areas

<table>
<thead>
<tr>
<th>Children:</th>
<th>CHILDES</th>
<th>PhonBank</th>
<th>Narrative</th>
<th>Bilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical:</td>
<td>AphasiaBank</td>
<td>FluencyBank</td>
<td>Dementia</td>
<td>TBIBank</td>
</tr>
<tr>
<td>Adult:</td>
<td>CABank</td>
<td>TutorBank</td>
<td>GestureBank</td>
<td>ClassBank</td>
</tr>
<tr>
<td>Multilingualism:</td>
<td>BilingBank</td>
<td>SLABank</td>
<td>Online Tutors</td>
<td>DOVE</td>
</tr>
</tbody>
</table>
## Funded Projects

<table>
<thead>
<tr>
<th></th>
<th>CHILDES</th>
<th>TalkBank</th>
<th>AphasiaBank</th>
<th>PhonBank</th>
<th>FluencyBank</th>
<th>LangBank</th>
<th>HomeBank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of Project (Years)</td>
<td>28</td>
<td>12</td>
<td>8</td>
<td>6</td>
<td>0.2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Words (millions)</td>
<td>59</td>
<td>47</td>
<td>1.8</td>
<td>0.8</td>
<td>0.5</td>
<td>2</td>
<td>audio</td>
</tr>
<tr>
<td>Linked Media (TB)</td>
<td>2.8</td>
<td>1.1</td>
<td>0.4</td>
<td>0.7</td>
<td>0.3</td>
<td>-</td>
<td>3.5</td>
</tr>
<tr>
<td># Languages</td>
<td>41</td>
<td>22</td>
<td>6</td>
<td>18</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td># Publications</td>
<td>7000</td>
<td>320</td>
<td>256</td>
<td>480</td>
<td>-</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td># Users</td>
<td>2950</td>
<td>930</td>
<td>390</td>
<td>182</td>
<td>25</td>
<td>-</td>
<td>22</td>
</tr>
<tr>
<td># Web Hits (millions)</td>
<td>4.1</td>
<td>1.3</td>
<td>0.3</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
41 languages (including Cantonese)

*MIC: 睇吓Sophie畫咗啲咩圖畫。
%mor: v | tai2=look_at asp | haa5=tentative n:prop | Sophie v | waak6=draw asp | zo2=perfective cl | di1=some wh | me1=what n | +n | tou4+n | waa2=drawing.

*CHI: 呢啲我哋彈琴架。
%mor: sfp | ne1=how_about cl | di1=some pro | ngo5-PL=I v | daan6=bounce n | kam4=piano sfp | gaa3.

*SIS: 你鍾意Alicia定係呀呀Lulu定係Sophie定係Timmy？
%mor: pro | nei5=you v | +v | zung1+n | ji3=like n:prop | Alicia conn | ding6hai6=or sfp | aa3 sfp | aa3 n:prop | Lulu conn | ding6hai6=or n:prop | Sophie conn | ding6hai6=or n:prop | Timmy ?
TalkBank Principles

❖ Community Driven
❖ Open access to data, media, derived corpora, and programs
❖ Standard format — CHAT, CHAT-XML, CHAT-CA
❖ CLAN programs running on CHAT format
❖ Transcripts linked to media
❖ Interoperable with other resources: R, Elan, Praat, SALT, Annis, CONLL, SpeechKitchen/Kaldi for ASR, LENA
❖ CHAT/PHON incorporates Praat
CLARIN Principles in TalkBank

❖ CLARIN-B center
  ❖ InCommon login through Shibbolet
  ❖ OAI-PMH server for OLAC, VLO
  ❖ DOI through HandleServer, EZ-Cite
❖ CLARIN-K center
  ❖ focus on analysis of spoken language
  ❖ tutorials for CLAN, video tutorials
  ❖ help desk, 5 Google Groups discussion boards
❖ SamtaleBank as a CLARIN illustration
Let’s take a look on the web

- childes.talkbank.org
- talkbank.org
- homebank.talkbank.org
- sla.talkbank.org
- talkbank.org/access/SamtaleBank
- childes.talkbank.org/browser — Alicia at 3;3
- downloadable materials
Major Methods

1. Corpus Analysis

2. Profiling

3. Microanalysis
1. Corpus Analysis

- FREQ - Frequency analysis
  - wild cards
  - word files (morality words, LIWC, medical)
- KWAL - Key word and line
  - matches highlighted
- COMBO - Regular expression matching
  - Hits can be triple-clicked to go back to transcript and play
LENA2CHAT

- 24 hour/day recordings in the home
- Much like Deb Roy’s database and the “water” example, but open
- Huge ITS files reduced automatically to manageable CHAT files
- Check out http://homebank.talkbank.org
Looking under the Hood
MOR, POST, GRASP

- 41 languages, but only 11 have MOR/POST
- Cantonese, Danish, Dutch, English, French, Italian, Hebrew, Japanese, German, Mandarin, Spanish
- GRASP for English, German, Hebrew, Spanish, Mandarin
MOR

- More declarative than FST
- Part-of-speech tuned to spoken language
- Easy to use once there is a grammar
- Hard to build the grammar (A-rules, C-rules)
- 98% accuracy for English
- POSTMORTEM rules for German declension
Bilingual MOR

- *CHL: +" [- spa] <yo no la> [/] yo no la desmentí porque. [+ break]
- *CHL: what's my word against hers &ladadada .
- *CHL: +" [- spa] todos estamos con un calor and@s working@s .
- All words are tagged implicity; can be made explicit.
- Coding system makes code-switching junctures evident.
- Run English MOR, excluding [- spa], then Spanish MOR including [- spa]
Dependency Graphs

Web service runs by triple-clicking on %gra line
2. Profiling - EVAL/KIDEVAL

- This all depends on MOR and GRASP
- Comparison database with s.d. scores
- IPSyn, DSS
- MLU, MLT
- TTR, vocD, MATTR
- Brown’s 14 morphemes
- TIMEDUR
EVAL

MLU, TTR
Verbs/Utt
% errors
% N, V, Aux, Adv, Conj, Pro
% PAST, PASTP, PL
Retracing, repetition
Sample Output
Analysis Pathways

Transcribe → A → MOR MEGRASP → KIDEVAL

CHATTER → B → PhonTalk → Phon → Praat
Error Analysis

- [*p] phonological p:w, p:n, p:m
- [* s] semantic s:r, s:ur. s:uk, s:per
- [* n] neologism n:k, n:uk, n:k:s, n:uk:s
- [* d] dysfluency
- [* m] morphology m:a:0es etc.
- [* f] formal lexical
- [+ gram] [+ jar] [+ es] [+ per] [+ cir]
3. Microanalysis (CA and Gesture)

Be: See what it said in here (. .) in- my theory (hhh) •
(0.4)

?: khu- (hhh)

Be: [about this amnesic (. .) dysnomic aphasia, •
(0.3)

Be: u:hm (it) says the cause of lesion is usually deep in
temporal lobe just like Kathy was saying presumably
interrupting connections of sensory speech areas with the
CHAT2ELAN
CHAT2PRAAT - sociophonetics

- Highlight utterance bullet
- Send to sound analyzer
- Extracts audio from video
- In Praat, draw a picture
CHAT2PHON
CHAT in ANNIS
CA Coding

- shift to high pitch; F1 up-arrow
- shift to low pitch; F1 down-arrow
- rising to high; F1 1
- rising to mid; F1 2
- level; F1 3
- falling to mid; F1 4
- falling to low; F1 5
- unmarked ending; F1 6
- continuation; F1 +
- inhalation; F1 .
- latching≈; F1 =
- uptake; F1 u
- top begin overlap; F1 [
- top end overlap; F1 ]
- bottom begin overlap; F1 { 
- bottom end overlap; F1 }

Δ - fasterΔ; F1 right-arrow

△ - slower△; F1 left-arrow

* - creaky*

/ - unsure?; F1 /

" - softer"; F1 0

© - louder©; F1 )

Δlow pitch_; F1 d

Δhigh pitch_; F1 h

Δsmile voice®; F1 l

® - breathy voice® marker; F1 b

© - whisper©; F1 w

¥ - yawn¥; F1 y

$ - singing$; F1 s

§ - precise§; F1 p

~ - constriction~; F1 n

pitch reset; F1 r

¬ - laugh in a word; F1 c

Tag or sentence final particle; F2 t

† - Vocative or summons; F2 v
Gestural Detail

- Interaction / Sequence / Segment
- Each participant coded through sequence
  - Deedee 1a-1b-1c
  - Nina 1a-1b-1c
- Bullet links each segment back to transcript
- Coding: gaze direction, action, classification, meaning, language
- Rapport coding through gaze, smile, language
Discourse Analysis

- CHAINS, KEYMAP, DIST
- CHIP
- PD (Propositional density)
- CI (Complexity index)
- SCRIPT + Speech Kitchen ASR
Time Series - corpora to R

Alberto and Jorge — I no go.
Collaborative Commentary
**Messages for CLARIN**

- Importance of open access
- Importance of uniform transcription format linked to analysis programs
- Importance of focus on specific research communities for:
  - corpus development
  - tool development
- FUNDING
Conclusions

- We need to expand TalkBank
- CLARIN can make wider use of TalkBank methods
- We can promote TalkBank-CLARIN integration